Executive Summary

The use of large-scale achievement tests as instruments of educational policy is growing. In particular, states and school districts are using such tests in making high-stakes decisions with important consequences for individual students. Three such high-stakes decisions involve tracking (assigning students to schools, programs, or classes based on their achievement levels), whether a student will be promoted to the next grade, and whether a student will receive a high school diploma. These policies enjoy widespread public support and are increasingly seen as a means of raising academic standards, holding educators and students accountable for meeting those standards, and boosting public confidence in the schools.

Because the stakes are high, the Congress wants to ensure that tests are used properly and fairly, and it asked the National Academy of Sciences, through its National Research Council, to "conduct a study and make written recommendations on appropriate methods, practices and safeguards to ensure that—

A. existing and new tests that are used to assess student performance are not used in a discriminatory manner or inappropriately for student promotion, tracking or graduation; and

B. existing and new tests adequately assess student reading and math-

ematics comprehension in the form most likely to yield accurate information regarding student achievement of reading and mathematics skills."

This study focuses on tests with high stakes for individual students. The committee recognizes that accountability for students is related in important ways to accountability for educators, schools, and school districts. Indeed, the use of tests for accountability of educators, schools, and school districts has significant consequences for individual students, for example, by changing the quality of instruction or affecting school management and budgets. Such indirect effects of large-scale assessment are worth studying in their own right. By focusing on the congressional interest in high-stakes decisions about individual students, this report does not address accountability at those other levels, apart from the issue of participation of all students in large-scale assessments.

BASIC PRINCIPLES OF TEST USE

The use of tests in decisions about student tracking, promotion, and graduation is intended to serve educational policy goals, such as setting high standards for student learning, raising student achievement-levels, ensuring equal educational opportunity, fostering parental involvement in student learning, and increasing public support for the schools. The committee recognizes that test use may have negative consequences for individual students even while serving important social or educational policy purposes. The development of a comprehensive testing policy should therefore be sensitive to the balance among the individual and collective benefits and costs of various uses of tests.

Determining whether high-stakes testing of students produces better overall educational outcomes requires that its potential benefits be weighed against its potential unintended negative consequences. Thus, the value of tests should also be weighed against the use of other information in making high-stakes decisions about students. Tracking, promotion, and graduation decisions will be made with or without tests.

The committee adopted three principal criteria, developed from earlier work by the National Research Council, for determining whether a test use is appropriate**:**

(1) measurement validity—whether a test is valid for a particular purpose, and whether it accurately measures the test taker's knowledge in the content area being tested;

(2) attribution of cause—whether a student's performance on a test reflects knowledge and skill based on appropriate instruction or is attributable to poor instruction or to such factors as language barriers or disabilities unrelated to the skills being tested; and

(3) effectiveness of treatment—whether test scores lead to placements and other consequences that are educationally beneficial.

These criteria, based on established professional standards, lead to the following basic principles of appropriate test use for educational decisions:

- The important thing about a test is not its validity in general, but its validity when used for a specific purpose. Thus, tests that are valid for influencing classroom practice, "leading" the curriculum, or holding schools accountable are not appropriate for making high-stakes decisions about individual student mastery unless the curriculum, the teaching, and the test(s) are aligned.
- Tests are not perfect. Test questions are a sample of possible questions that could be asked in a given area. Moreover, a test score is not an exact measure of a student's knowledge or skills. A student's score can be expected to vary across different versions of a test—within a margin of error determined by the reliability of the test—as a function of the particular sample of questions asked and/or transitory factors, such as the student's health on the day of the test. Thus, no single test score can be considered a definitive measure of a student's knowledge.
- An educational decision that will have a major impact on a test taker should not be made solely or automatically on the basis of a single test score. Other relevant information about the student's knowledge and skills should also be taken into account.
- Neither a test score nor any other kind of information can justify a bad decision. Research shows that students are typically hurt by simple retention and repetition of a grade in school without remedial and other instructional support services. In the absence of effective services for low-performing students, better tests will not lead to better educational outcomes.

The committee has considered how these principles apply to the appropriate use of tests in decisions about tracking, promotion, and graduation, to increasing the participation of students with disabilities and English-language learners in large-scale assessments, and to possible uses

of the proposed voluntary national tests in making high-stakes decisions about individual students. The committee has also examined existing and potential strategies for promoting appropriate test use.

USES AND MISUSES OF TESTS

Blanket criticisms of tests are not justified. When tests are used in ways that meet relevant psychometric, legal, and educational standards, students' scores provide important information that, combined with information from other sources, can lead to decisions that promote student learning and equality of opportunity. For example, tests can identify learning differences among students that the education system needs to address. Because decisions about tracking, promotion, and graduation will be made with or without testing, proposed alternatives to the use of test scores should be at least equally accurate, efficient, and fair.

It is also a mistake to accept observed test scores as either infallible or immutable. When test use is inappropriate, especially in making high-stakes decisions about individuals, it can undermine the quality of education and equality of opportunity. For example, the lower achievement test scores of racial and ethnic minorities and students from low-income families reflect persistent inequalities in American society and its schools, not inalterable realities about those groups of students. The improper use of test scores can reinforce these inequalities. This lends special urgency to the requirement that test use with high-stakes consequences for individual students be appropriate and fair.

Decisions about tracking, promotion, and graduation differ from one another in important ways. They differ most importantly in the role that mastery of past material and readiness for new material play. Thus, the committee has considered the role of large-scale high-stakes testing in relation to each type of decision separately in this report. But tracking, promotion, and graduation decisions also share common features that pertain both to appropriate test use and to their educational and social consequences.

Members of some minority groups, English-language learners, and students from low socioeconomic backgrounds are overrepresented in lower-track classes and among those denied promotion or graduation on the basis of test scores. Moreover, these same groups of students are underrepresented in high-track classes, "exam" schools, and "gifted and talented" programs. In some cases, such as courses for English-language

learners, such disproportions are logical: one would not expect to find native English speakers in classes designed to teach English to English-language learners. In other circumstances, such disproportions raise serious questions. For example, grade retardation among children cumulates rapidly after age 6, and it occurs disproportionately among males and minority group members. These disproportions are especially disturbing in view of other evidence that, as typically practiced, grade retention and assignment to low tracks have little educational value. For example, assignment to low tracks is typically associated with an impoverished curriculum, poor teaching, and low expectations. It is also important to note that group differences in test performance do not necessarily indicate problems in a test, because test scores may reflect real differences in achievement. These, in turn, may be due to a lack of access to a high-quality curriculum and instruction. Thus, a finding of group differences calls for a careful effort to determine their cause.

RECOMMENDATIONS

The committee offers more detailed recommendations in Chapter 12 about the appropriate uses of tests. Those recommendations cover cross-cutting issues that affect testing generally; specific issues and problems pertaining to the uses of tests in tracking, promotion, and graduation; and the inclusion of students with disabilities and students who are English-language learners. The organization of the recommendations in Chapter 12 follows the logic of the chapters in this report. In this executive summary, we present overarching recommendations and discuss the possible use of the proposed voluntary national tests for high-stakes decisions about individual students.

- Accountability for educational outcomes should be a shared responsibility of states, school districts, public officials, educators, parents, and students. High standards cannot be established and maintained merely by imposing them on students. Moreover, if parents, educators, public officials, and others who share responsibility for educational outcomes are to discharge their responsibility effectively, they should have access to information about the nature and interpretation of tests and test scores. Such information should be freely available to the public and should be incorporated into teacher education and into educational programs for principals, administrators, public officials, and others.

• Tests should be used for high-stakes decisions about individual mastery only after implementing changes in teaching and curriculum that ensure that students have been taught the knowledge and skills on which they will be tested. Some school systems are already doing this by planning a gap of several years between the introduction of new tests and the attachment of high stakes to individual student performance, during which schools may achieve the necessary alignment among tests, curriculum, and instruction. But others may see attaching high stakes to individual student test scores as a way of leading curricular reform, not recognizing the danger that such uses of tests may lack the "instructional validity" required by law—that is, a close correspondence between test content and instructional content.

• The consequences of high-stakes testing for individual students are often posed as either-or propositions, but this need not be the case. For example, "social promotion" and repetition of a grade are really only two of many educational strategies available to educators when test scores and other information indicate that students are experiencing serious academic difficulty. But neither social promotion nor retention alone is an effective treatment for low achievement, and schools can use a number of other possible strategies to reduce the need for these either-or choices, for example, by coupling early identification of such students with effective remedial education.

• Some large-scale assessments are used to make high-stakes decisions about individual students, but most often in combination with other information, as recommended by the major professional and scientific organizations concerned with testing. For example, most school districts say they base promotion decisions on a combination of grades, achievement test scores, developmental factors, attendance, and teacher recommendations. As our study has shown, however, a number of jurisdictions have adopted policies that rely exclusively on achievement test scores to make high-stakes decisions. A test score, like other sources of information, is not exact. It is an estimate of the student's understanding or mastery at a particular time. Therefore, high-stakes educational decisions should not be made solely or automatically on the basis of a single test score but should also take other relevant information into account.

• The preparation of students plays a key role in appropriate test use. It is not proper to expose students ahead of time to items that will actually be used on their test or to give students the answers to those questions. Test results may also be invalidated by teaching so narrowly to the objectives of a particular test that scores are raised without actually im-

proving the broader set of academic skills that the test is intended to measure. The desirability of "teaching to the test" is affected by test design. For example, it is entirely appropriate to prepare students by covering all the objectives of a test that represents the full range of the intended curriculum. We therefore recommend that test users respect the distinction between genuine remedial education and teaching narrowly to the specific content of a test. At the same time, all students should receive sufficient preparation for the specific test so their performance will not be adversely affected by unfamiliarity with its format or by ignorance of appropriate test-taking strategies.

- Accurate assessment of students with disabilities and English-language learners presents complex technical and policy challenges, in part because these students are particularly vulnerable to potential negative consequences when high-stakes decisions are based on tests. We recommend that policymakers pursue two key policy objectives in modifying tests and testing procedures in these special populations:

(1) to increase such students' participation in large-scale assessments, in part so that school systems can be held accountable for their educational progress; and

(2) to test each such student in a manner that provides appropriate accommodation for the effect of a disability or of limited English proficiency on the subject matter being tested, while maintaining the validity and comparability of test results among all students.

These objectives are sometimes in tension, and the goals of full participation and valid measurement thus present serious technical and operational challenges to test developers and users.

- The purpose of the proposed voluntary national tests (VNT) is to inform students (and their parents and teachers) about their performance in 4th grade reading and 8th grade mathematics relative to the standards of the National Assessment of Educational Progress and to performance in the Third International Mathematics and Science Study. The proposal does not suggest any direct use of VNT scores to make decisions about the tracking, promotion, or graduation of individual students, and thus it is not being developed to support those uses. However, states and school districts would be free to use scores on the voluntary national tests for these purposes. Given their design, the proposed voluntary national tests should not be used for decisions about the tracking, promotion, or

graduation of individual students. The committee takes no position on whether the voluntary national tests are practical or appropriate for their primary stated purposes.

- The committee sees a strong need for better evidence on the intended benefits and unintended negative consequences of using high-stakes tests to make decisions about individuals. A key question is whether the consequences of a particular test use are educationally beneficial for students—for example, by increasing academic achievement or reducing dropout rates. It is also important to develop statistical reporting systems of key indicators that will track both intended effects (such as higher test scores) and other effects (such as changes in dropout or special education referral rates). Indicator systems could include measures such as retention rates, special education identification rates, rates of exclusion from assessment programs, number and type of accommodations, high school completion credentials, dropout rates, and indicators of access to high-quality curriculum and instruction.

PROMOTING APPROPRIATE TEST USE

At present, professional norms and legal action (through administrative enforcement or litigation) are the principal mechanisms available to enforce appropriate test use. These mechanisms are inadequate. Compliance with provisions of the Joint Standards for Educational and Psychological Testing and the Code of Fair Testing Practices in Education is largely voluntary, and enforcement is often weak. Legal action is typically adversarial, time-consuming, and expensive, and applicable law can vary by jurisdiction, making enforcement uneven.

New methods, practices, and safeguards could take any of several forms, but in general they would appear at various points on a continuum between professional norms and legal enforcement, some less coercive, some more so. Deliberative forums, an independent oversight body, labeling, and federal regulation represent a range of possible options that could supplement professional standards and litigation as means of promoting and enforcing appropriate test use.

The committee is not recommending adoption of any particular strategy or combination of strategies, nor does it suggest that these four approaches are the only possibilities. We do think, however, that ensuring proper test use will require multiple strategies. Given the inadequacy of current methods, practices, and safeguards, there should be further re-

search on these and other policy options to illuminate their possible effects on test use. In particular, we would suggest empirical research on the effects of these strategies, individually and in combination, on testing products and practice, and an examination of the associated potential benefits and risks.

Large-scale assessments, used properly, can improve teaching, learning, and equality of educational opportunity. That tests are sometimes used improperly should not discourage policymakers, teachers, and parents. Rather, it should motivate action to ensure that educational tests are used fairly and effectively. This report is a contribution to that essential work.